Decision Trees for Hierarchical Classification of Transposable Elements

Bruna Zamith Santos BRUNA.ZAMITH@HOTMAIL.COM
Federal University of São Carlos, Department of Computer Science - São Paulo - Brazil

Rafael Gomes Mantovani RGMANTOV@ICMC.USP.BR
University of São Paulo, Department of Computer Science - São Paulo - Brazil

Leander Schietgat LEANDER.SCHIETGAT@CS.KULEUVEN.BE
KU Leuven, Department of Computer Science, Leuven, Belgium

Celine Vens CELINE.VENS@KULEUVEN-KULAK.BE
KU Leuven Kulak, Department of Public Health and Primary Care, Kortrijk, Belgium

Ricardo Cerri CERRI@DC.UFSCAR.BR
Federal University of São Carlos, Department of Computer Science - São Paulo - Brazil

Keywords: transposable elements, hierarchical classification, machine learning, decision trees

Abstract

Transposable Elements (TEs) are DNA sequences that can change their location within the genome. Accurate classification of TEs is an important step towards understanding their effects on genes and their role in genome evolution. Usually, TEs classification is performed using homology-based tools, comparing a sequence with a database with many sequences belonging to previously known TE classes. This is a limited strategy, since it ignores the sequences' biochemical properties, and also the hierarchical relationships that may exist between the different TE classes. Based on the existing proposals to establish a hierarchical TE taxonomy, we propose a preliminary study of TE classification methods using Machine Learning (ML). The ML methods will then be compared with existing literature methods, and evaluated using measures specifically designed for hierarchical classification problems.

Appearing in *Proceedings of Benelearn 2016*.

1. Introduction

Transposable Elements (TEs) are DNA sequences that can move and duplicate within genomes, autonomously or with assistance of other elements. Accurate TEs classification enables research into their biology and shed light on the evolutionary processes that shape genomes (Wheeler et al., 2013).

TEs in eukaryotes can be classified according to whether reverse transcription is needed for their transposition (Class I or retrotransposons) or not (Class II or DNA transposons). Although a consensus for a universal TE classification has not been reached yet (Piégu et al., 2015), there are some attempts to establish a hierarchical taxonomy of TEs. The hierarchical system proposed by (Wicker et al., 2007) is among the most accepted (Figure 1). It includes the levels of class, subclass, order, superfamily, family and subfamily. Class I is composed of five orders: LTR retrotransposons, DIRS-like elements, Penelope-like elements (PLEs), long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs).

A widely used method for TE classification is RepeatMasker (Smit et al., 2010). This tool is used to find and mask repeats in query sequences according to their similarity with sequences from a given annotated library. TEClass (Abrusn et al., 2009) classifies sequences with a hierarchy of binary classifiers based

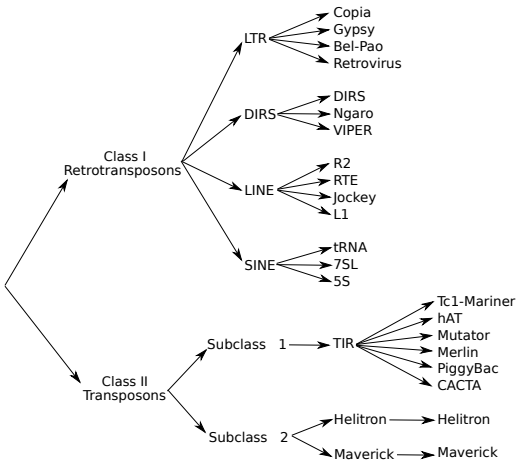

Figure 1. TE hierarchy as introduced by (Wicker et al., 2007).

on machine learning support vector machines, using oligomer frequencies as features. Recently, PASTEC (Hoede et al., 2014) was proposed. It also uses several features of TEs to classify TE sequences: structural features (TE length, presence of a LTR or TIR, presence of simple sequence repeats, etc.), sequence similarities to known TEs, and conserved functional domains found in HMM profile databases. Importantly, none of the existing classification systems is able to provide classifications at the superfamily level, and also none of them consider the hierarchical relationships between classes.

Since TE classes can be hierarchically structured, we will investigate TE classification as a ML hierarchical classification problem (HC) Silla2010, trying to take the hierarchical class relationships into consideration, and working with the entire Wicker07:jrnl taxonomy, which has yet not been done in the literature. HC is a very challenging problem in ML, since hundreds or even thousands of classes are involved, and the prediction task becomes more difficult as we traverse the hierarchy towards the more specific leaf classes. Thus, one of our contributions is to construct new TE datasets for the ML community.

2. Methods

To construct the ML datasets, we downloaded TE sequences from two public databases, Repbase (Jurka et al., 2005) and PSGB Plant database (Nussbaumer et al., 2013). Based on previous works (Costa et al., 2007; Costa et al., 2008), we wanted to check if protein profile signatures could be used as good features for TE classification. As not all TE sequences are translated to amino acid sequences, we used the Interproscan tool (Jones et al., 2014) to find all open reading frames (ORFs) associated to each DNA sequences. Having these ORFs, we could then use the ps-scan tool (de Castro, 2006) to match them against the PROSITE (Sigrist et al., 2002) collection of signatures, which represent sequence patterns. The presence/absence of these signatures was them used as features for ML algorithms. At total, we collected 430 features (PROSITE signatures). The use of PROSITE signatures as features was already performed for protein function prediction (Costa et al., 2008), but never in the context of TEs classification. Figure 2 shows the pipeline proposed for the construction of our TE dataset.

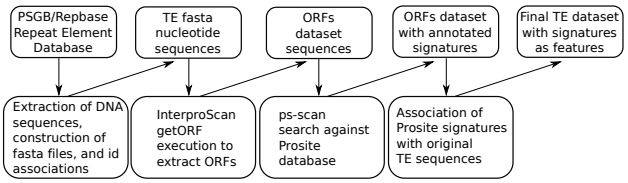

Figure 2. Pipeline for the TE dataset construction.

As a preliminary study, we evaluated the performance of Decision Tree induction algorithms considering only the leaf nodes of Wicker's hierarchy. We used the Quintan's Decision Tree induction algorithm (Quinlan, 1993) implemented in RWeka (Hornik et al., 2009), and considered the following TE categories in the experiments: Copia, Gypsy, tRNA, 7SL, 5S, Tc1-M (Mariner), hAT and Mutator.

3. Results and Final Considerations

Table 1 shows the error matrix generated by the Decision Tree after a 10-fold cross-validation experiment. As can be observed, we could get better results in the more frequent classes of the dataset, obtaining an accuracy of 79.68%. These are specific classes located at the leaf nodes of the hierarchy (Gypsy and Tc1-M), which may be an indicative that the features we are using (PROSITE signatures) may be very discriminative in some cases. Also, these classes are located at the superfamily level, where none of the existing methods had provided classification.

We consider that the poor results in some classes come from the class unbalance and also the hierarchy itself. This preliminary dataset we constructed has very few positive instances for some classes, for example Copia and 5S, and many positive instances for other classes, for example Gypsy, all located in a same hierarchical level. Also, hierarchical methods could discriminate classes in a top-down fashion, which is advantageous,

Table 1. Decision Tree Error Matrix

	Gypsy	Tc1-M	tRNA	hAt	Copia	Mutator	7SL	5S	Total
Gypsy	16089	459	34	162	0	0	3	0	16747
Tc1-M	221	1810	69	238	0	1	3	0	2342
tRNA	29	145	160	84	0	0	37	0	455
hAt	287	941	112	669	0	4	23	0	2036
Copia	1553	0	0	0	0	0	0	0	1553
Mutator	111	113	26	45	0	21	4	0	320
7SL	2	5	47	2	0	0	39	0	95
5S	0	16	4	6	0	0	3	0	29
Total	18292	3489	452	1206	0	26	112	0	23577

considering that we could build specialized classifiers for each class.

We plan to add more ML algorithms to our experiments, and to propose classifiers that take hierarchical relationships while inducing the model. These classifiers could take advantage of the fact that classes closer to the root have more positive instances than more deeper classes, which could improve the results. We also plan to investigate other types of attributes, which can be more discriminative.

Acknowledgments

We acknowledge CNPq and FAPESP for financial support, specially grant #2015/14300-1 from São Paulo Research Foundation (FAPESP).

References

Abrusn, G., Grundmann, N., DeMester, L., & Makalowski, W. (2009). Teclass - a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics*, *25*, 1329–1330.

Costa, E. P., Lorena, A. C., Carvalho, A. C., & Freitas, A. A. (2008). Top-down hierarchical ensembles of classifiers for predicting g-protein-coupled-receptor functions. *III Brazilian Symposium on Bioinformatics* (pp. 35–46). Berlin, Heidelberg: Springer-Verlag.

Costa, E. P., Lorena, A. C., Carvalho, A. C., Freitas, A. A., & Holden, N. (2007). Comparing several approaches for hierarchical classification of proteins with decision trees. *II Brazilian Symposium on Bioinformatics* (pp. 126–137). Berlin, Heidelberg: Springer-Verlag.

de Castro, L. N. (2006). *Fundamentals of Natural Computing: Basic Concepts, Algorithms, and Applications*. Chapman & Hall/CRC.

Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., & Quesneville, H. (2014). Pastec: An automatic transposable element classification tool. *PLoS ONE*, *9*, e91929.

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning: R meets Weka. *Computational Statistics*, *24*, 225–232.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). Interproscan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)*, *30*, 1236–1240.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*, *110*, 462–467.

Nussbaumer, T., Martis, M. M., Roessner, S. K., Pfeifer, M., Bader, K. C., Sharma, S., Gundlach, H., & Spannagl, M. (2013). Mips plantsdb: a database framework for comparative plant genome research. *Nucleic Acids Research*, *41*, D1144–D1151.

Piégu, B., Bire, S., Arensburger, P., & Bigot, Y. (2015). A survey of transposable element classification systems - a call for a fundamental update to meet the challenge of their diversity and complexity. *Mol Phylogenet Evol.*, *86*, 90–109.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Sigrist, C. J. A., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., & Bucher, P. (2002). Prosite: A documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics*, *3*, 265–274.

Smit, A., Hubley, R., & Green, P. (2010). Repeatmasker open-3.0. www.repeatmasker.org.

Wheeler, T., Clements, J., Eddy, S., Hubley, R., Jones, T., Jurka, J., Smit, A., & Finn, R. (2013). Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic acids research*, *41*, D70–D82.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., San Miguel, P., & Schulman, A. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*, 973–982.